Genre-adaptive Semantic Computing and Audio-based Modelling for Music Mood Annotation

Pasi Saari, György Fazekas, *Member, IEEE,* Tuomas Eerola, Mathieu Barthet, *Member, IEEE,* Olivier Lartillot, and Mark Sandler, *Senior Member, IEEE,*

Abstract—This study investigates whether taking genre into account is beneficial for automatic music mood annotation in terms of core affects valence, arousal, and tension, as well as several other mood scales. Novel techniques employing genre-adaptive semantic computing and audio-based modelling are proposed. A technique called the ACTwg employs genre-adaptive semantic computing of mood-related social tags, whereas ACTwg-SLPwg combines semantic computing and audio-based modelling, both in a genre-adaptive manner. The proposed techniques are experimentally evaluated at predicting listener ratings related to a set of 600 popular music tracks spanning multiple genres. The results show that ACTwg outperforms a semantic computing technique that does not exploit genre information, and ACTwg-SLPwg outperforms conventional techniques and other genre-adaptive alternatives. In particular, improvements in the prediction rates are obtained for the valence dimension which is typically the most challenging core affect dimension for audio-based annotation. The specificity of genre categories is not crucial for the performance of ACTwg-SLPwg. The study also presents analytical insights into inferring a concise tag-based genre representation for genre-adaptive music mood analysis.

Index Terms—Music information retrieval, mood prediction, social tags, semantic computing, music genre, genre-adaptive.

1 Introduction

Musical genre and mood are closely linked together. People tend to use particular genres for mood balancing [1]. Different genres are able to induce distinct emotional responses [2] while mood and genre terms are often combined to express musical qualities (e.g. "smooth jazz" and "dark ambient") [3]. In the field of Music Information Retrieval (MIR), automatic music annotation and retrieval in terms of moods and genres have received considerable attention [4], [5], [6], [7]. Moreover semantic metadata related to mood and genre have been shown to be amongst the most important ones for machine-based semantic music annotation or auto-tagging [8], [9]. It is easy to see why. On one hand, psychological studies have shown that music can be organised according to perceived and induced (elicited) emotions[1] [2], [10], and music's ability to convey and affect moods is a key factor in explaining why music is culturally important [11]. On the other hand, music genres have traditionally been the most common music content descriptors aiming to categorise music for sales, delivery and consumption (e.g. in retail stores, radio, libraries) [6]. Music genres account for a majority of social tags – free-form textual labels or phrases collaboratively applied to particular resources by users – in online music services such as Last.fm[2] [12].

Automatic mood annotation or mood prediction of modern online music catalogues that span tens of millions of tracks from numerous genres in a semantically meaningful manner requires advanced computational techniques. Benefits of audio-based techniques relying on features related to rhythm, timbre, tonality and others have been shown in numerous music mood annotation studies [4], [13], [14], [15]. In particular, these techniques are beneficial since they solve the cold-start problem [16] of music indexing, providing labels for music not yet rated by people. However, audio-based techniques rely on human-generated ground-truth at the model training stage. Generating such ground-truth for large data sets in a controlled manner is often prohibitively laborious [17]. This may lead to a bottleneck for reaching successful model performance, since the size of the available data may not be sufficiently representative of larger and more heterogeneous music collections.

Other ground-truth sources, more abundant but arguably less reliable, have been exploited to deal with the issue of limited data availability. These sources can be divided into the games-with-a-purpose [18], [19], online editorial tags [20], [21] and social tags [22], [23]. Of these, social tags provide the most extensive resource for semantic information on music, but their free-form nature leads to problems related to subjective error and noise, synonymy, polysemy,

- *P. Saari is with the Department of Music, University of Jyvaskyla, 40014 Jyvaskyla, Finland. E-mail: pasi.saari@jyu.fi*
- *G. Fazekas, M. Barthet and M. Sandler are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. E-mail: g.fazekas@qmul.ac.uk; m.barthet@qmul.ac.uk; mark.sandler@qmul.ac.uk*
- *T. Eerola is with the Department of Music, Durham University, DH1 3RL Durham, U.K. E-mail: tuomas.eerola@durham.ac.uk*
- *O. Lartillot is with the Department for Architecture, Design and Media Technology, Aalborg University, DK-9000 Aalborg, Denmark. E-mail: olartillot@gmail.com.*

Manuscript received Xxxx 00, 0000; revised Xxxxxxx 00, 0000.

1. We employ the words emotion and mood interchangeably in the present paper.

2. http://www.last.fm

and data sparsity [24]. Semantic computing techniques have been employed successfully to tackle these problems [9]. For the purpose of mood annotation, these techniques have been applied in a bottom-up manner to learn emotion models in line with those suggested by research in affective sciences [25], [26]. These learnt models have been deemed efficient as semantic representations [22], [27] and proved robust at smoothing out the noise prevalent in tag data [23]. Audio-based techniques have been applied successfully in conjunction with tag-based semantic computing techniques, either by treating computational audio features as "quasi-tags" alongside textual tags [9], or by mapping the audio features to tag-based semantic layers [20], [28], [29], [30]. The benefit of the latter family of techniques is that they require only the audio file at the prediction stage.

Using large data sources for music mood annotation has two advantages: 1) it enables operating on data that better represent large modern-day music catalogues and thus enables tapping into global characteristics of the relationship between music and emotion; and simultaneously, 2) it enables drawing information on this relationship at a more detailed level, by considering genre-specific aspects for example. While audio-based techniques have been efficient at predicting genres [31], mood prediction has remained more elusive [32], especially on the valence dimension, relating to the distinction between positive and negative emotions [14], [30], [33]. Taking into account the genre-specificity of music moods may provide a way to alleviate this issue. Certain mood tags are more relevant to one genre than to another [3] and audio-based mood annotation models trained on sets of tracks drawn from a particular genre give more accurate predictions within the corresponding genres than across genres [34].

Two general approaches have been proposed for audio-based genre-adaptive mood annotation: the *genre-feature approach* and the *genre-split/combine approach*. The genre-feature approach treats genre tags as conventional input features, either on their own, or alongside a set of audio features [35]. The genre-split/combine approach involves splitting training data into genre subsets, training multiple mood prediction models on these subsets, and when annotating a novel music item, combining the outputs of each model according to the genre of the item. Genre of the novel item may be determined either by pre-specified labels or by using a separate audio-based genre annotation model. In past research, techniques employing the genre-split/combine approach have outperformed equivalent non genre-adaptive techniques [36], [37], whereas techniques employing the genre-feature approach provided a similar level of performance as techniques employing audio features only [35].

To summarise, employing audio-based techniques, semantic computing and genre-adaptivity in music mood annotation have been beneficial in recent studies, while results have been boosted by using semantic computing or genre-adaptivity in conjunction with audio-based techniques. However, it has not yet been investigated whether mood prediction performance could be boosted further by *genre-adaptive semantic computing* or by combining genre-adaptive semantic computing with audio-based techniques. The present study offers the following novel contributions for music mood annotation: 1) it proposes a technique that employs genre-adaptive semantic computing; 2) it proposes a technique that employs both semantic computing and audio-based mood annotation in a genre-adaptive manner; and 3) it assesses the effect of the specificity of genre categories for genre-adaptive mood annotation. The benefit of the genre-adaptive techniques is evaluated against a number of baseline techniques including non genre-adaptive techniques, conventional auto-tagging and the genre-feature approach. The techniques are trained on a large set of social tag data and audio, and evaluated for the prediction of listeners' ratings of the perceived moods in a separate set of music tracks.

The rest of the paper is organised as follows: Section 2 discusses how this work relates to the previous studies in MIR. Section 3 describes the data covered in the study while Sections 4 and 5 delineate the techniques to represent and annotate music in terms of mood and genre. Section 6 introduces genre-adaptive mood prediction techniques, and finally, Sections 7 and 8 report the results and conclude the paper.

2 Related Work

The majority of research in music and mood has utilised either categorical (e.g., happiness, sadness and anger) [26] or dimensional models of emotion. A well-known example of the latter is the affective circumplex [25] which represents different emotions in the underlying dimensions of valence, distinguishing between positive and negative emotions, and arousal, relating to the activity or the intensity aspect of emotion. These dimensions, as well as the tension dimension, spanning from relaxed to tense emotions, have been described as *core affects* [38], [39]. However, valence and arousal have also been considered as the primary dimensions, while tension has been inferred as the product of negative valence and positive arousal [40]. Both the categorical and the dimensional model have been employed in audio-based music mood annotation for classifying music into discrete mood categories [15], [41], [42], or for predicting the core affect and other mood dimensions using regression models [14], [42], [43].

Semantic computing techniques, often based on Latent Semantic Analysis (LSA) [44] have been employed to represent music moods based on tag data. In particular, a model resembling the affective circumplex, inferred in a bottom-up manner from tag data, has been found robust at representing the mood of music [22], [23], [27]. Other representations such as the categorical model have been investigated as well [22]. Saari & Eerola [23] proposed an LSA-based technique called the Affective Circumplex Transformation (ACT) that yielded significant improvements over other techniques, as well as raw tags, at predicting listener ratings of the perceived core affects in music. The model training was carried out using social tags from Last.fm and a follow-up study confirmed the results using curated editorial tags associated to production music tracks [21]. The Semantic Layer Projection technique (SLP) was proposed in [29] as an extension to ACT to enhance audio-based music mood prediction. SLP involves projecting tracks to moods via a two-stage process, whereby a corpus of tracks is first mapped based on associated tags to a semantic space

obtained with ACT, and then multiple regression models are trained between audio features and the semantic space. SLP outperformed conventional regression models trained to map audio features directly to listener ratings [29], [30]. ACT and SLP are employed as building blocks of the novel genre-adaptive techniques introduced in the present study.

Several studies have highlighted the challenges of representing the genre of music. In particular, the finding optimal "resolution" of genres [45] and the fuzziness of genre categories [46] are important problems. In studies attempting to identify the underlying factors of music preferences based on genres, four [47] and five [48] underlying factors have typically been singled out. In other studies, 10 [7], 13 [49], and 16 [35] genres have been employed to characterise the typical diversity of music. The analysis of artist tags retrieved from Last.fm highlighted the fuzziness of genre categories [46]: for instance, 56% of artists tagged with "pop" were also tagged with "rock", and 87% of "alternative" music overlapped with "rock". Still these three genres have been considered as separate categories in typical music catalogues such as iTunes. The evidence from social tags indicates that a single genre describing a track is not inclusive enough, but perhaps a (weighted) combination of several genre labels would better describe genre information.

Previous approaches to music auto-tagging have gained performance improvements by taking into account the relationships between tags, as observed in their co-occurrence patterns or correlations [50], [51], [52]. For example, Ness et al. [50] trained Support Vector Machine (SVM) models with probabilistic outputs first for multiple tags separately and then used the outputs of each tag-specific model as inputs to second-stage SVMs, which enabled taking into account the relationships between tags. These types of techniques have outperformed non-contextual models and in particular, stacked SVMs have yielded state-of-the-art performance [3]. However, these techniques exploit tag relationships irrespective of whether the tags relate to genres, moods or other concepts.

For audio-based mood annotation, considering genre in particular as contextual information has led to positive results. Lin et al. [37] employed the genre-split/combine approach to the auto-tagging of music in terms of moods, training multiple genre-specific models and combining the models at the annotation stage. They used album-level editorial tags from the Allmusic.com service to represent each music track. Compared to a general model, their genre-adaptive model increased the F-score performance from 0.23 to 0.36. Similar results were obtained in [36] for the classification of music to mood clusters. In comparison with these studies, the present study employes genre-adaptive semantic computing in addition to audio-based modelling and predicts moods represented by dimensions rather than categories or binary classes. Rather than using album-level editorial tags, each track in the test data is rated by several dozens of participants. Moreover, the difference between the nature of the training and test ground-truth, i.e. large-scale but unreliable social tags and reliable listener ratings of mood dimensions, arguably improves the ecological validity of the results thus obtained.

3. cf. http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results

TABLE 1
The employed data sets.

	$ntracks$	Modalities	$nterms$ per track Mood	Genre
TR100k	118,874	tags	3.23	5.38
TR10k	10,199	tags, audio	3.56	5.52
TE600	600	tags, audio, ratings	7.22	8.53

3 Data Collection

This section introduces the data comprising mood- and genre-related social tags and associated audio tracks. Table 1 summarises the statistics of the employed data sets, of which TR100k and TR10k are used for model training, and TE600 is reserved for performance evaluation.

3.1 Training Data Sets

The social tag data collected from Last.fm in [23] and reused this study consists of 924k unique tags associated with 1.3M tracks. Each track-tag association is represented by weights in $\mathbb{Z}_{[0-100]}$. As in [23], mood- and genre-related tags were identified by string matching against large lists of mood and genre terms gathered from various sources. For moods, each tag that included a term as a substring was linked to the corresponding term[4], whereas for genres, the tags which were kept were only those that fully matched one of the terms. The resulting set was further reduced by keeping only the first 100 mood terms and 100 genre terms that were associated to the highest number of tracks. Tracks performed by artists appearing in TE600, and tracks that were not associated to any mood or any genre term were then excluded.

TR10k, including audio for 10,199 tracks, was sampled from the set resulting from the process above. Full-length CD quality audio files were obtained by accessing the I Like Music (ILM) catalogue, a curated music database with accurate metadata. Last.fm tracks were paired with ILM tracks using controlled track sampling method based on several potentially conflicting criteria. The aim was to ensure a close match between Last.fm and ILM track by using low Levenshtein string distance between the metadata entries (artist, track and album names) with less than 0.5s difference between track durations. The number of tracks within each expert-generated genre category available from ILM was balanced to ensure a fair coverage of different genres overall. The maximum number of tracks sampled from the same artist was limited to avoid artist and album effects. Finally, a three-dimensional mood space obtained by ACT in [23] was used as basis to provide a good coverage of the mood space for each genre. The tracks were sampled such that their distribution in this space is as close to uniform as possible. The resulting TR10k dataset includes tracks from 5,470 unique artists.

4. E.g., tag "happy mood" was thus linked to the term "happy". If several tags of a track matched the same mood term the highest weight was used.

TR100k was formed by augmenting TR10k with all of the initial corpus that were performed by any artists in TR10k. This was necessary because track sampling excluded important semantic information present in the original set of tracks. This resulted in a set of 118,847 tracks. As seen in Table 1, the average number of terms associated to each track is higher for genres than for moods, even after the exact string matching of genre terms. This obviously reflects the overall higher prevalence of genre tags than mood tags in social tag data reported in [46]. Within TR10k (and TR100k shown in parentheses), a median of 162 (1,687) tracks are associated to a mood term and a median of 329 (3,869) to a genre term. The most prevalent mood terms are "chillout" (2,569) and "party" (1,638), whereas the least prevalent mood terms are "pleasant" (51) and "bliss" (51). The most prevalent genre terms are "rock" (3,587) and "pop" (3,091), whereas the least prevalent are "root reggae" (147) and "jazz fusion" (150). The relative term prevalences are roughly the same within TR100k.

For the experimental evaluations, 10 training partitions, each comprising 80% of tracks in TR100k or TR10k were randomly subsampled from the training data. The subsequent evaluations were thus carried out by performing the model training separately on each training partition and applying the resulting models on the full TE600 set. We will denote the partitions as T (within TR100k) and T' (within TR10k), $T' \subset T$.

3.2 Test Data Set

The TE600 reserved for the evaluation is the same as the one described in [23]. The set consists of 600 tracks with Last.fm tags, audio files and listener ratings of perceived moods. Six broad genres including Metal, Rock, Folk, Jazz, Electronic and Pop are represented in this dataset. The mood ratings were collected from 59 participants on nine step Likert scales for all core affects (Valence, Arousal and Tension) and seven mood terms (Atmospheric, Happy, Dark, Sad, Angry, Sensual, and Sentimental). The ratings were summarised by the average across participants. This deemed sufficient due to the high consistency between the participants reported in [23]). Although listener ratings were obtained for 15 second clips, we use the full tracks in the present study, relying on the claim made in [23] that the clips are representative of the full tracks. The ratings and links to the audio and tag data is publicly available[5].

The tag data associated to TE600 was subjected to a similar process applied to the training sets: each track was linked to the 100 mood and 100 genre terms selected for the training data, and tracks not associated to any mood or genre term were excluded (12 in total).

3.3 Audio Features

62 audio features related to dynamics, onsets, autocorrelation, chromagram, and spectrum were extracted from the full-length tracks of TR10k and TE600 using the MIRtoolbox[6] [53]. These are summarised in Table 2. The audio material was first summed to mono and cut into overlapping analysis frames with feature-specific lengths and degrees of overlap. A frame length of 50ms with 50% overlap was used for low-level spectral features, MFCCs and their first (Δ) and second order ($\Delta\Delta$) instantaneous derivatives and for all features related to dynamics. Audio onsets were detected from temporal amplitude curves extracted from a 10-channel filter bank decomposition. Event density was calculated by the number of onsets in 10s, 50% overlapping frames. The features derived from the autocorrelation were calculated using 3s frames with 90% overlap (33.3% overlap for Tempo). Finally chromagrams were computed using 750ms, 50% overlapping frames. From this, several high-level features related to tonality were calculated such as Mode (majorness) and Key clarity.

TABLE 2
Audio features, aggregated to $*$: *mm*, *ms*, *sm* and *ss*; †: *mm* and *sm*. The "Frame" column reports the window lengths and overlaps ($\star$: 50ms length with 50% overlap).

Category	Feature	Stats	Frame
Dynamics	RMS, Zero-crossing rate	$*$	$\star$
Onsets	Attack (time, slope, leap)	$*$	Onset-based
	Event density	†	10s, 50%
Autocorrelation	Pulse clarity, Novelty	†	3s, 90%
	Tempo	†	3s, 33.3%
Chromagram	Mode, HCDF, Key Clarity, Centroid, Novelty	†	750ms, 50%
Spectrum	Novelty, Brightness, Centroid, Spread, Flux, Skewness, Entropy, Flatness, Roughness	$*$	$\star$
	13 coef. MFCC, Δ, $\Delta\Delta$	$*$	$\star$

All features with different frame lengths where brought to the same time granularity by computing the Mean (*m*) and standard deviation (*s*) over 1s, 50% overlapping texture windows. However, only the Mean was computed for Event density and in case of the chromagram-related features because they were extracted from longer frames to begin with. Similarly, the standard deviations were omitted for the MFCC derivatives, since their mean values already describe the temporal change. Finally, 178 song-level descriptors were obtained by taking again the Mean (*mm* and *ms*) and Standard deviation (*sm* and *ss*) over the texture window frames. This process is motivated by the approach presented in [50]. Typically the song-level representation of audio features is calculated as the Mean and Standard deviation over the whole track length, excluding the texture window processing. The approach taken here incorporates the temporal dynamics of the features at both short and long time span in a more sensitive fashion compared to the typical song-level averaging approach.

4 General techniques for Semantic Computing and Audio-based Annotation

Conventional techniques that do not take genre information into account are explained in this section.

5. http://hdl.handle.net/1902.1/21618
6. MIRtoolbox version 1.5.

4.1 Semantic Computing Using ACT

First, ACT [23] was applied on the training partitions of TR100k to enable representing the mood of tracks based on the associated tags. Initially, associations between mood terms i and tracks j are represented in a standard Vector Space Model (VSM) matrix $M = m_{i,j}$. As in [23], M was first normalised by computing Term Frequency-Inverse Document Frequency (TF-IDF) scores and then transformed to a three dimensional semantic mood space by applying non-metric Multi-Dimensional Scaling (MDS). Note, that in [23], dimension reduction was employed in two stages by applying the Singular Value Decomposition (SVD) prior to the MDS. This provided a slight performance improvement compared to dimension reduction with MDS only. However, since the number of mood terms is lower in the present study (100 compared to 357), the SVD stage was excluded. This allowed a reduction in the number of alternative model parameterisations (e.g., the number of dimensions in SVD) in the experiment. In the next stage, a Valence-Arousal space (VA space) was inferred by conforming the semantic mood space to a reference configuration of mood terms (cf. below). This was done using the Procrustes transformation [54] that performs a linear mapping from a space to another while retaining relative distances between objects in the original space. This yields a configuration $X_i = (x_{i,1}, x_{i,2}, x_{i,3})$ of all mood terms in the VA space, retaining the third dimension as in [23].

To apply ACT for mood annotation, a track j represented by mood VSM vector $q_{i,j}$ was projected to the VA space by first normalizing the vector according to the learned TF-IDF scores yielding $\hat{q}$. Then, a representation $S_j = (s_{j,1}, s_{j,2}, s_{j,3})$ of the track in the VA space was computed by

$$S_j = \frac{\sum_i \hat{q}_i x_i}{\sum_i \hat{q}_i}. \quad (1)$$

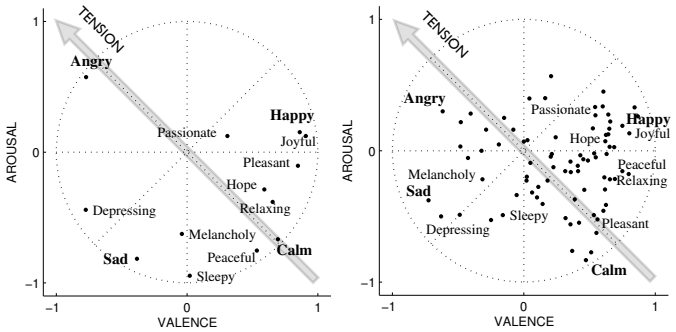
The final estimates $P_j^{(a)}$ related to the core affects and mood terms were obtained by

$$P_j^{(a)} = \frac{a}{|a|} \cdot S_j, \quad (2)$$

where a equals to the term positions in the configuration X_i for each mood term i, and $(1, 0, 0)$, $(0, 1, 0)$, and $(-1, 1, 0)$ for Valence, Arousal and Tension respectively.

Of the 101 mood terms present in Russell's and Scherer's reference configuration [25], [55], 13 terms could be matched with the 100 mood terms used in the present study. This configuration, plotted onto the VA space in Fig. 1a, is denoted by *Russell*. As one can see, most of the matched terms are located in the low Arousal – high Valence quadrant. Due to this imbalance, a more simple reference configuration was formed by including only one mood term for each VA quadrant: Happy, Calm, Sad, and Angry indicated in boldface in Fig. 1a. This configuration is denoted *Russell4*. These terms were chosen since they are frequently cited in music and emotion research [56] and their prevalence within TR100k was above the median (*10,459*, *3,554*, *9,306* and *1,921* for Happy, Calm, Sad, and Angry respectively).

Affective norm data related to a large set of English lemmas [57] was also explored as a direct alternative to the mood term positions inferred using the ACT. To use this configuration, the MDS and Procrustes stages in the ACT were skipped. In addition to Valence and Arousal, the data includes Dominance as the third dimension. 81 mood terms, summarised in Fig. 1b, could be matched between the norm data and tags. This configuration is denoted *Norms*. To train the model with *Norms*, 339 tracks had to be excluded from TR10k since they were not associated to any of the matched mood terms.



(a) Russell (**Russell4**) (b) Norms

Fig. 1. Reference mood term configurations from a) Russell [25] and Scherer [55]; and b) Affective norms [57].

4.2 Audio-based Annotation Using the SLP

SLP involves training a set of regression models to map audio features to the VA space dimensions learnt using ACT, and applying these models to predict moods in music tracks. In [29] and [30] Partial Least-Squares (PLS) was employed as a regression technique for SLP, whereas in the present study the LIBSVM implementation of Support Vector Regression (SVR) [58] was used. This allowed a direct comparison to an SVM auto-tagger (cf. Section 4.3).

The audio features related to TR10k were z-score-transformed to a zero mean and unit standard deviation. Extreme values were considered outliers and truncated to $[-5, 5]$. To reduce the SVR training time, highly correlated audio features were removed using agglomerative hierarchical clustering with the correlation distance function. To this end, the complete linkage criterion with a cutoff correlation distance of 0.1 was employed and the first feature in each obtained cluster according to the order presented in Table 2 was kept.

As in Section 4, the VA space was learned using ACT and tracks in the TR10k were projected to the VA space based on the associated tags. SVR models were then trained to map the pre-processed audio feature set to each of the VA space dimensions separately. In a preliminary analysis the SVR was tested using the linear and Radial Basis Function (RBF) kernels, but results indicated that the linear kernel gives a performance comparable to the RBF with a shorter training time. The cost parameter c was set to 0.001 since it yielded consistently high performance compared several candidates $c = 10^y, y = [-4, -3, ..., 1]$. SLP was applied on the test data to produce audio-based estimates S'_j. Finally, estimates $P_j^{(a)'}$ related to the core affects and mood terms were computed, similar to those described in Section 4.1.

4.3 Audio-based Annotation using SVM Auto-tagger

Two-stage stacked SVMs [50] were employed to compare the SLP performance to a conventional auto-tagger using an implementation following [50]. First, the input audio features of TR10k were pre-processed as described in Section 4.2 and the mood tags were transformed into binary classes. The first-stage SVM classifiers with a linear kernel and probabilistic outputs were trained separately for each mood and the classifiers were applied on the training tracks. The obtained positive class probabilities for all moods were then served as input to the second-stage SVM classifiers to again map the input to the binary mood classes. When annotating a track in the test data, the models were applied to produce a vector of probability estimates. This vector was normalised to sum to one. Note, that the stacked SVMs are not capable of directly producing estimates for core affects, since Valence, Arousal and Tension are not explicitly represented by any of the mood terms.

Since the mood term prevalence in TR10k varies from 0.5% to 26%, the binary tag data fed to the SVMs is highly imbalanced. In the past, taking into account the class imbalance has yielded positive results for SVM-based music mood auto-tagging [37]. Therefore cost-sensitive learning found effective in [59] was employed by setting different misclassification error costs for the positive and negative class related to each mood. The costs $c_i^+ = 1$ and $c_i^- = n_i^+/n_i^-$ were set for the positive and negative classes respectively (n_i^+ and n_i^- are the number of the positive and negative tracks within the training data for a mood i).

To form another baseline technique, tracks were projected to the VA space based on the outputs of the stacked SVMs. The outputs were TF-IDF-weighted and projected to the VA space as in the ACT prediction stage. This technique, as opposed to the original stacked SVM, is inherently capable of producing estimates also for the core affects. Similar baseline techniques were implemented already in [30] using PLS regression to predict the normalised tag counts, but using stacked SVMs instead proved more efficient as the results will show. These two baseline techniques are denoted *SVM-orig* and *SVM-ACT*.

5 Genre Clustering and Representation Based on Tags

Prior to exploiting genres as contexts in mood annotation, a sufficiently concise genre representation was sought after. This was done to reduce the computational burden of the mood annotation techniques and because the majority of distinct genres might be too narrow in terms of mood content for within-genre semantic analysis of moods. To this end, genre term clustering was applied to reduce the number of distinct genres.

5.1 Genre Clustering Techniques

Given the associations between genre term i and track j in a VSM matrix $G = g_{i,j}$, the rows g_i were grouped into disjoint clusters $C = \{C_1, C_2, ..., C_K\}$ using the following techniques and specifications:

K-means: G was first normalised to a unit Euclidean length by

$$\hat{g}_{i,j} = g_{i,j}/(\sum_{j \in T} {g_{i,j}}^2)^{1/2}, \quad (3)$$

after which the algorithm was run using the cosine distance.

Agglomerative hierarchical clustering: G was first normalised according to the TF-IDF to produce $\hat{G}$. The cosine distance between $\hat{g}_i$ and $\hat{g}_{i'}$ was then used as the distance measure, and the agglomeration was done based on the average link criterion.

Spectral clustering: G was normalised according to the TF-IDF, and Cosine similarities between $\hat{g}_i$ and $\hat{g}_{i'}$ were used as the affinity matrix. Clustering was then done following the method described in [60], similar to [36] where the technique was applied to group emotion tags.

5.2 Genre Clustering Survey and Evaluation

To assess the quality of the obtained genre clusterings, an online genre grouping survey was organised. This also provided insight into the number of genre clusters that would optimally represent the data. The survey task for each participant was to arrange the 100 genre terms into any number of clusters between 2-16 they considered most appropriate. The range for the candidate number of genres was selected to acknowledge the typical number of genres assessed in past studies. The instructions specified that the clusters should group genres that share common musical, social or cultural characteristics. The participants were asked to be objective in their assignments, and the instructions allowed using any external web resource to check the definition of possible unfamiliar genre terms (e.g., "downtempo"). 19 participants, predominantly engineering and musicology students knowledgeable of different music genres took part in the survey.

No clear optimal number of genre clusters arose from the survey results. The number of clusters ranged between 6 and 16, with peaks around 9, 10 and 16 clusters (M = 11.34, SD = 3.17). To validate this result, the conventional Davies-Bouldin technique [61] was applied on the genre tag data which allows to infer the optimal number of clusters. This analysis did not yield a clear optimum either. This may reflect the general difficulty of defining the genre granularity that would satisfy all purposes.

Genre clusterings were computed based on the training partitions of TR100k using $K = \{2, 4, 6, ..., 16\}$ clusters. The clusterings were compared to those obtained from the survey using the Mirkin metric [62], which can be used to assess the disagreement between two clusterings $C = \{C_1, C_2, ..., C_K\}$ and $C' = \{C'_1, C'_2, ..., C'_{K'}\}$ by:

$$d_M(C, C') = \sum_k n_k^2 + \sum_{k'} n_{k'}^2 - 2 \sum_k \sum_{k'} n_{kk'}^2, \quad (4)$$

where n and n_k are the numbers of genre terms in G and cluster C_k, respectively and $n_{kk'}$ is the number of terms in $C_k \bigcap C'_{k'}$. This metric can be used to compare clusterings with different K. For identical clusterings, $d_M = 0$, and $d_M > 0$ otherwise. The d_M values, computed separately between the tag-based clusterings and each of the 19 survey clusterings, were averaged across the participants and training partitions. The results are shown in Fig. 2. One can see

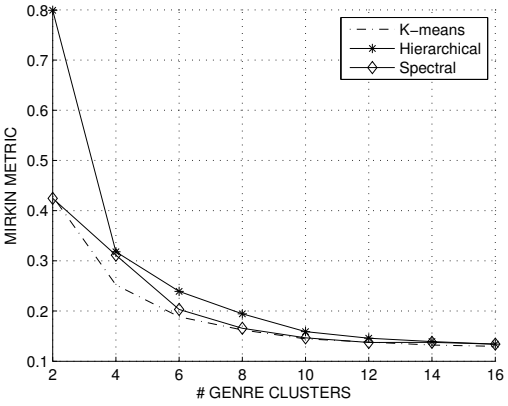


Fig. 2. The average Mirkin metric of each genre clustering technique.

TABLE 3
Genre clusters obtained using K-means with $K = \{2, 4, 6, ..., 16\}$.

K	Most prevalent genre term
2	Pop, Rock
4	Soul, Rock, Hard rock, Electronic
6	Hard rock, Singer songwriter, Electronic, Jazz, Rock, Pop
8	Electronic, Rnb, Soul, Instrumental, Pop, Singer songwriter, Rock, Hard rock
10	Soul, Hip hop, Rock, Electronic, Singer songwriter, Reggae, Alternative, Jazz, Metal, Lounge
12	Electronic, Downtempo, Country, Soul, Hard rock, Punk, Rnb, Singer songwriter, Rock, Jazz, Classic rock, Pop
14	Hip hop, Rock, Singer songwriter, Pop, Pop rock, Jazz, Country, Soul, Metal, New wave, Hard rock, Classic rock, Instrumental, Electronic
16	Jazz, Rnb, Instrumental, Reggae, Ambient, Pop rock, Rock n roll, Experimental, New wave, Classic rock, Pop, Soul, Electronic, Hard rock, Singer songwriter, Rock

that all clustering techniques compare similarly to the survey data, except that the hierarchical clustering performed poorly with $K = 2$. In general, K-means outperformed the other techniques by a slight margin. Therefore K-means was used in subsequent analyses.

In order to examine the genre clustering results in more detail, clusterings were computed using $K = \{2, 4, ..., 16\}$ based on the full TR100k set. Table 3 shows the most prevalent genre tag for the genre clusters obtained in this manner.

Although the survey did not give clear indication of the optimal number of genre clusters, subsequent analyses were primarily conducted with $K = 6$. This was the minimum number obtained in the survey. It also corresponds well with the six broad genre categories of TE600. Fig. 3 shows in detail the discrepancy between the clustering with $K = 6$ and the survey data. For each pair of genre terms the number of participants that assigned both terms to the same cluster was computed. Six genre terms most prevalent in the TR100k are shown for each cluster in the order of prevalence. The six clusters correspond well with the main genres in TE600 since each of these terms are in different clusters. Therefore the clusters are labeled with these genres. One can see from the figure that genre terms in Metal, Folk and Electronic were mostly grouped together also by

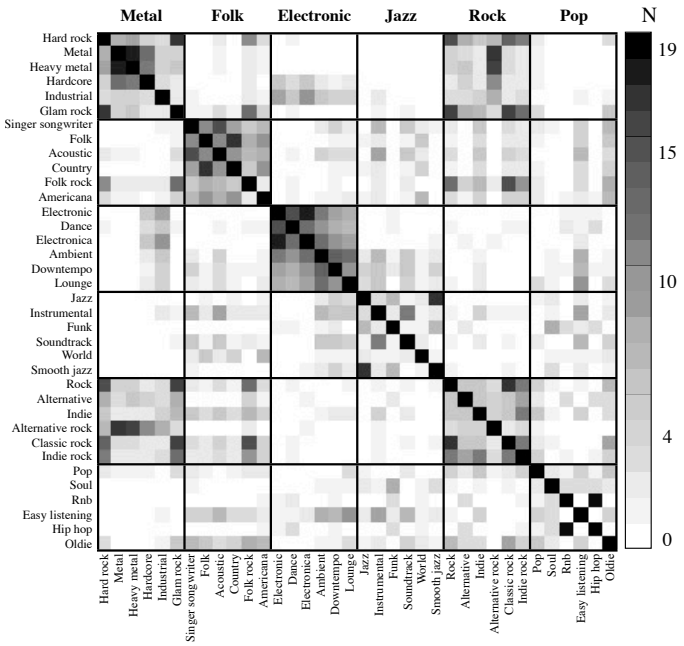


Fig. 3. The discrepancy between the six genre clusters obtained using the K-means, and the cluster co-occurrences of genre terms obtained from the survey.

TABLE 4
The percentage of tracks in the data sets associated to each of the six genre clusters.

Genre cluster	TR100k	TR10k	TE600
Metal	18.2	14.0	27.0
Folk	28.7	34.0	40.3
Electronic	30.3	31.8	46.2
Jazz	33.6	41.3	44.0
Rock	60.1	55.3	80.2
Pop	49.2	57.1	66.2

participants, whereas terms in Jazz, Rock and Pop were not grouped as consistently.

5.3 Track-level Genre Representation

Given the associated genre tags and a genre clustering C, the genre of a track j was represented by a weighted combination $H = h_{k,j}$ ($k \in \{1, 2, ..., K\}$) of the associated genre clusters:

$$h_{k,j} = \frac{\sum_{i \in C_k} \hat{g}_{i,j}}{n_k} \left[\sum_k \frac{\sum_{i \in C_k} \hat{g}_{i,j}}{n_k} \right]^{-1}, \quad (5)$$

where $\hat{g}_{i,j}$ was computed with (3) based on the full TR100k set. Table 4 shows the percentage of tracks in the data sets that are positively associated to each genre cluster. One can see that the clusters are very broad: 80.2% and 66.2% of TE600 tracks belong to Rock and Pop respectively. The high prevalence of tags related to Pop and Rock reflects the fuzziness of these genres.

5.4 Audio-based Genre Annotation using the SVM Auto-tagger

For the audio-based genre annotation stacked SVMs were trained on the genre tags similar to *SVM-orig*. Given the audio features of a novel track, the vector of genre term probabilities were first predicted, and then the vector was mapped with (5) to the genre clusters obtained using K-means.

6 Genre-Adaptive Mood Annotation

The genre-adaptive techniques, employing either the genre-feature or the genre-split/combine approach, incorporate ACT for semantic computing and SLP for audio-based annotation. When the techniques are used to annotate novel tracks, two variants are applied: one predicting genres from tag data and one predicting genres from audio.

It the training phase, these techniques use mood and genre tag data as input: audio features A and a clustering of genre terms $C = \{C_1, C_2, ..., C_K\}$. The tag-based genre representation $h_{k,j}$ is then computed with (5). In the prediction phase, tag-based genre representation is computed again with (5), whereas audio-based genre representation is computed as described in Section 5.4.

6.1 Genre-feature Techniques

The genre-feature techniques use genre information as normal input features for SLP training:

Genre-based Prediction (SLPg): *SLPg* involves predicting mood using only the genre information as input giving an indication as to how much variance of mood ratings can be attributed to mood prevalence differences between genres. *SLPg* differs from the general SLP in that the input features are represented by $h_{k,j}$ instead of audio features – all other stages are the same as in general SLP.

Genre- and Audio-based Prediction (SLPga): *SLPga* is similar to *SLPg*, but uses the audio features alongside the genre information. The data is therefore represented by $(h_{1,j}, h_{2,j}, ..., h_{K,j}, a_{1,j}, a_{2,j}, ..., a_{n,j})$, where n is the number of features after pre-processing.

6.2 Genre-Split/Combine Techniques

All genre-split/combine techniques involve splitting training data into (possibly overlapping) genre subsets, for either ACT training (denoted by ACTwg), SLP training (denoted by SLPwg), or both (denoted by ACTwg-SLPwg). The ACT training is performed within TR100k, whereas SLP training is performed within TR10k. Splitting is done based on the genre tags so that a subset related to genre k comprises the tracks $\{j : h_{k,j} > 0\}$. Fig. 4 shows how the techniques differ from the general form of ACT and SLP.

Genre-adaptive Semantic Computing (ACTwg): ACTwg is based on the assumptions that relationships of moods vary between genres and that genre-specific semantic models are required to boost the mood annotation performance of semantic computing. The final model combining these genre-specific models would then sufficiently account for the variation between genres. To train the ACTwg model, K mood term configurations X_i^k in genre-specific VA spaces are learned using ACT and at the prediction stage, these

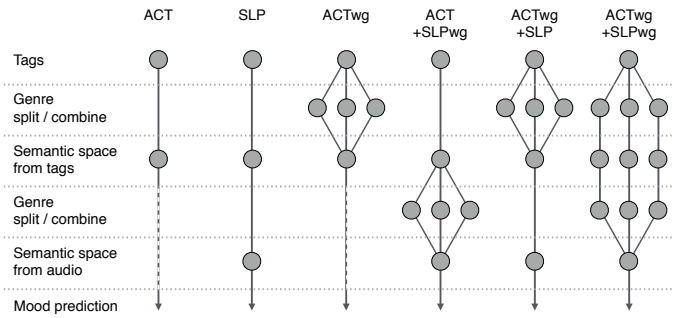


Fig. 4. A schematic diagram showing the different stages at which genre-adaptivity is applied in the genre-split/combine techniques.

models are applied to produce the genre-specific estimates $P_j^{(a_k)_k}$. The final estimates are then computed by weighting the genre-specific estimates proportionately to h_k:

$$P_j^{(a)} = \frac{1}{\sum_k h_{k,j}} \sum_k h_{k,j} P_j^{(a_k)_k}. \quad (6)$$

Audio-based modelling within Genres (ACT-SLPwg): In ACT-SLPwg, the general type of semantic computing is performed and only the audio-based SLP models are trained within each genre subset. The assumption underlying this technique is that audio features and feature combinations relate to moods differently within different genres. SLP models are trained on each genre subset as described in Section 4.2. Applying these models on novel tracks produces genre-specific estimates $P_j^{(a)'_k}$. The final estimates are then computed similar to (6):

$$P_j^{(a)'} = \frac{1}{\sum_k h_{k,j}} \sum_k h_{k,j} P_j^{(a)'_k}. \quad (7)$$

Genre-adaptive Semantic computing and audio-based modelling (ACTwg-SLP): In ACTwg-SLP, it is assumed that genre-adaptive semantic computing is needed but that the relationship between audio and VA space dimensions remains static across genres. First, genre-specific ACT models are trained similarly to ACTwg, and the models are applied on the tracks in the training data to produce the estimates for the mood dimensions:

$$S_j = \frac{1}{\sum_k h_{k,j}} \sum_k h_{k,j} S_j^k. \quad (8)$$

Then, general SLP models are trained to map the audio to S_j. At the prediction stage, the general SLP models are applied to produce $P_j^{(a_k)'}$ and the final estimates are computed by

$$P_j^{(a)'} = \frac{1}{\sum_k h_{k,j}} \sum_k h_{k,j} P_j^{(a_k)'}. \quad (9)$$

Genre-adaptive Semantic Computing and Genre-adaptive Audio-based modelling (ACTwg-SLPwg): ACTwg-SLPwg employs genre-adaptivity in both semantic computing and audio-based modelling, assuming that both the semantic relationships of mood terms and audio-to-mood associations vary between genres. First, genre-specific ACT models are trained on each genre subset and the models are applied on the training data to produce S_j^k.

Then, genre-specific SLP models are trained to map audio to S_j^k. At the prediction stage, the final estimates are computed by

$$P_j^{(a)'} = \frac{1}{\sum_k h_{k,j}} \sum_k h_{k,j} P_j^{(a_k)'_k}. \quad (10)$$

7 Results and Discussion

The annotation performance of the techniques was evaluated in terms of the coefficient of determination statistic (R^2) after fitting simple linear regression models between the estimates and the listener ratings. The R^2-statistic was chosen as the goodness-of-fit measure since it is used in the bulk of past studies on automatic prediction of Arousal and Valence for music[7]. Median and median absolute deviation (MAD) across the models trained on each of the training partitions of TR100k and TR10k are reported.

7.1 General Techniques

7.1.1 Tag-based Annotation

First, the performance of the ACT models trained using the different mood term configurations was compared so as to choose the most successful configuration for subsequent analyses. The results are shown in Table 5. In general, the core affects were more easy to predict than the mood terms, and the performance for Valence was lower than for Arousal. These findings are in line with those obtained from past work evaluating ACT with TE600 data [23], [30]. *Russell4* yielded the highest performance for seven mood scales, and was clearly more efficient than *Russell* for Dark and Sad. These mood terms were also among the most difficult to predict. On the other hand, *Russell* was more successful at predicting Valence, Tension and Atmospheric. *Norms* yielded dramatically lower performance than the other configurations, which arguably supports exploiting music-specific data to form the semantic mood space, rather than using a mood configuration that relates to affective connotations of mood words in general. It also indicates that the inclusion of Dominance as the explicit third dimension in the mood space does not provide clear benefits. When examining the average performance across mood scales, *Russell4* ($R^2 = 0.387$) outperformed *Russell* by a slight margin ($R^2 = 0.371$). This suggests that ACT is not overly sensitive to changes in the mood reference configuration and that a simple reference configuration provides a strong enough reference to reliably represent mood terms in the VA space. Therefore, *Russell4* was chosen for the subsequent audio-based analyses.

The VA space obtained using *Russell4* is presented in Fig. 5. The mood positions for the figure were computed as the average of those obtained from each training partition. The underlying dimensions of Valence and Arousal are easily distinguishable, and the obtained positions for the four reference terms correspond fairly well with the original positions, with the exception of Sad, which is close to neutral in the Valence dimension. This finding is in line with [10], where musical examples expressing sadness were perceived as neutral in terms of Valence.

7. This statistic equals to the squared Pearson's correlations and was chosen since the mood estimates, roughly within $[-1.5, 1.5]$, are scaled differently to the ratings, which were given on scales from 1 to 9.

TABLE 5
Prediction results for ACT with *Russell* and *Russell4* reference configurations and *Norms*.

	Russell	*Russell4*	*Norms*
Valence	**0.425** 0.013	0.413 0.015	0.270 0.000
Arousal	0.477 0.004	**0.486** 0.003	0.269 0.000
Tension	**0.382** 0.015	0.378 0.014	0.219 0.001
Atmospheric	**0.424** 0.039	0.395 0.026	0.157 0.000
Happy	0.384 0.016	**0.386** 0.011	0.300 0.000
Dark	0.274 0.087	**0.348** 0.035	0.038 0.000
Sad	0.201 0.015	**0.276** 0.013	0.166 0.000
Angry	0.522 0.013	**0.531** 0.017	0.214 0.000
Sensual	0.403 0.006	**0.416** 0.007	0.002 0.000
Sentimental	0.220 0.020	**0.238** 0.023	0.061 0.000
Average	0.371	0.387	0.170

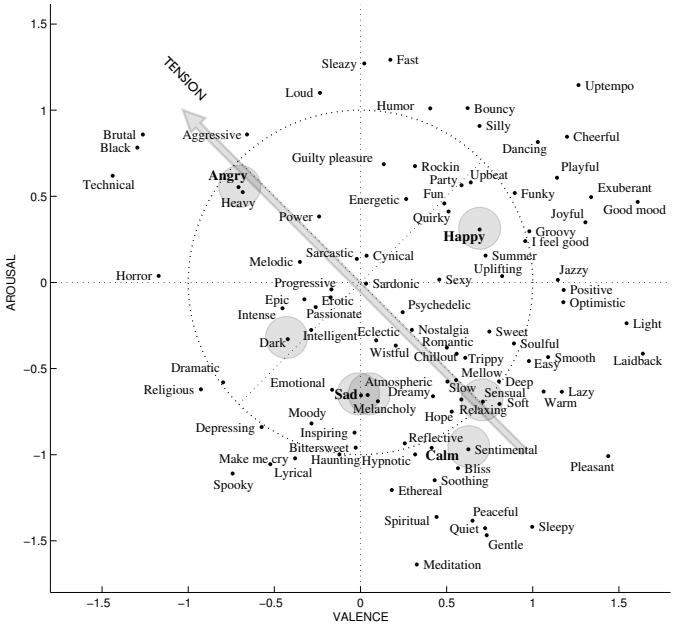


Fig. 5. Mood tag positions (the averages across training partitions) obtained with ACT using *Russell4* as the reference configuration.

7.1.2 Audio-based Annotation

Table 6 presents the performance obtained with SLP (using *Russell4*) and stacked SVMs. The audio-based mapping onto the VA-space using SLP provided dramatically higher performance than the tag-based mapping using ACT for all mood scales except for Valence, Happy, and Dark – all of which in fact relate to either positive or negative moods. The clearest difference between the SLP and the ACT was obtained for Arousal ($R^2 = 0.728$ vs. 0.477). This rather surprising result, although congruent with that reported in [30], may be explained by the sparsity and the inherent unreliability of tag data: the ACT maps tracks to the mood space based on only few tags, which may cause local inconsistencies. By contrast, mapping audio features to the mood dimensions using SLP may tap into more global patterns and provide a way to "smooth out" these inconsistencies. The mean SLP performance across mood scales was similar to that reported in [30] ($R^2 = 0.455$ vs. 0.453). However, prediction performance for Valence was clearly higher in the present study, $R^2 = 0.359$ compared to $R^2 = 0.322$.

SVM-orig performed inconsistently, whereas *SVM-ACT*

TABLE 6
Prediction results for SLP and SVM baseline techniques.

	SLP	SVM-orig	SVM-ACT
Valence	0.359 0.019	–	**0.369** 0.020
Arousal	**0.728** 0.004	–	0.714 0.005
Tension	**0.485** 0.019	–	0.483 0.022
Atmospheric	**0.696** 0.014	0.069 0.004	0.684 0.020
Happy	0.312 0.030	0.205 0.003	**0.314** 0.031
Dark	0.235 0.023	**0.311** 0.004	0.248 0.020
Sad	0.303 0.011	0.316 0.007	**0.323** 0.007
Angry	0.589 0.016	**0.622** 0.008	0.618 0.017
Sensual	**0.544** 0.004	0.252 0.026	0.535 0.010
Sentimental	0.300 0.024	**0.436** 0.016	0.304 0.030
Average	0.455	0.316	0.459

TABLE 7
Genre prediction performance in terms of median and MAD across training partitions.

	Precision	Recall	AP	AROC
Metal	0.776 0.010	0.569 0.003	0.826 0.004	0.841 0.002
Folk	0.642 0.010	0.531 0.015	0.734 0.006	0.755 0.003
Electronic	0.800 0.010	0.515 0.009	0.852 0.005	0.769 0.002
Jazz	0.698 0.013	0.577 0.004	0.795 0.006	0.766 0.003
Rock	0.918 0.004	0.524 0.003	0.939 0.001	0.731 0.001
Pop	0.850 0.004	0.629 0.011	0.888 0.004	0.781 0.003

(employing *Russell4*) increased the performance by a clear margin. Low performance of *SVM-orig* for Atmospheric, Happy, and Sensual suggests that the way these tags are applied by Last.fm users is not accounted well by musical characteristics, and that the musical characteristics congruent with these mood dimensions are better modelled by more general patterns incorporated in a low-dimensional mood space. Although *SVM-ACT* provided performance comparable to SLP, the benefit of SLP is lower computationally complexity requiring one audio-based model for each VA space dimension. Therefore, SLP may be considered the best-performing technique to be used as the baseline for genre-adaptive techniques.

7.2 Audio-based Genre Prediction

Performance of audio-based genre prediction was assessed by comparing the predicted values to the tag data. Although the reliability of social tags is questionable, the tag-based evaluation was considered sufficient for the present study because of the subsidiary role of audio-based genre prediction. Table 7 shows the performance for each genre cluster in terms of the standard evaluation metrics Precision, Recall, Average Precision (AP) and the area under the ROC curve (AROC). For each track, the SVM produces probability estimates related to the association strength of each genre cluster. To compute the Precision, Recall and AP, three genres with the highest probability were considered as positive for each track. AROC, on the other hand, was computed based on all probability values[8]. The results showed that genre prediction from audio is sufficient (see [8] for comparison) and may be used as an alternative to tag-based genre inference.

8. See [8] for detailed explanation of these metrics

TABLE 8
Performance of the genre-feature techniques with genres inferred from tags and audio. Performance improvements over the SLP are highlighted.

	Tag-based genres		Audio-based genres	
	SLPg	SLPga	SLPg	SLPga
Valence	**0.372** 0.003†	**0.453** 0.008 *	0.346 0.003†	**0.400** 0.020† *
Arousal	0.146 0.004† *	0.702 0.004† *	0.234 0.004† *	**0.731** 0.005
Tension	0.278 0.007† *	0.463 0.012† *	0.349 0.009† *	**0.494** 0.018†
Atmosph.	0.205 0.016† *	0.662 0.025 *	0.294 0.013† *	**0.704** 0.021
Happy	0.221 0.003† *	**0.398** 0.020 *	0.175 0.002† *	**0.332** 0.034†
Dark	**0.379** 0.011† *	**0.378** 0.025† *	**0.326** 0.016† *	**0.271** 0.019
Sad	0.000 0.000† *	0.271 0.023† *	0.009 0.002† *	0.296 0.009†
Angry	0.501 0.004† *	**0.647** 0.009 *	0.571 0.008† *	**0.630** 0.019 *
Sensual	0.341 0.015† *	0.532 0.026	0.379 0.008† *	**0.546** 0.009
Sentim.	0.058 0.003† *	0.246 0.021† *	0.096 0.003† *	0.296 0.028†
Average	0.250	0.475	0.278	0.470

† $p < .05$ for performance difference between the ACTwg-SLPwg.
* $p < .05$ for performance difference between the SLP.

7.3 Genre-adaptive Techniques

To assess the statistical significance of performance differences between the general and the genre-adaptive techniques, Wilcoxon rank sum tests were carried out across models trained on the training partitions. The techniques involving audio-based mood prediction were compared to general SLP (see Table 6), whereas ACTwg was compared to the general ACT (see *Russell4* in Table 5). Furthermore, equivalent comparisons were carried out between each technique and ACTwg-SLPwg. Results for the genre-feature and the genre-split/combine techniques are presented in Tables 8 and 9, respectively.

7.3.1 Genre-feature Approach

Among the techniques employing the genre-feature approach, SLPga performed well compared to general SLP. With audio-based genres, it outperformed the general model for all mood scales except for Sad and Sentimental. It yielded clear improvements of the average performance across mood scales. On the other hand, already the SLPg, relying only on genres as inputs, yielded relatively high performance for Valence, the most challenging core affect for general SLP. Moreover, with the genre-split/combine techniques included, SLPg was interestingly the most successful technique for Dark, which indicates that Dark correlates highly with genre information.

7.3.2 Genre-split/combine Approach

Results for ACTwg showed that genre-adaptive semantic computing is beneficial for tag-based mood annotation. ACTwg outperformed general ACT for all mood scales except for Atmospheric and Sensual. This performance difference was significant for five scales. Performance improvement over the ACT was the most notable for Valence with ACTwg reaching $R^2 = 0.457$.

Genre adaptivity of audio-based modelling using ACT-SLPwg improved the performance especially for Valence, Happy, Dark, and Angry, which suggests that the musical characteristics correlating with moods related to the positive/negative emotions differ between genres. With the exception of Angry, these mood scales were also among

the most difficult to predict using the general SLP. ACTwg-SLP also improved the performance over general SLP and yielded performance comparable to that of ACT-SLPwg. In general however, these techniques were not significantly more effective than the more simple genre-feature technique SLPga.

ACTwg-SLPwg yielded the highest performance. The average performance across moods was $R^2 = 0.482$ with tag-based genres and $R^2 = 0.492$ with audio-based genres. These figures are considerably higher than that of the general SLP ($R^2 = 0.455$). Notably, ACTwg-SLPwg with audio-based genres gave statistically significant improvements over SLP for seven mood scales and importantly for all core affects. The clearest performance improvement was achieved for Valence, where the ACTwg-SLPwg yielded $R^2 = 0.457$ with tag-based genres and $R^2 = 0.431$ with audio-based genres. Also for Arousal and Tension, ACTwg-SLPwg with audio-based genres yielded the highest performance of $R^2 = 0.741, 0.520$ respectively. The fact that audio-based genre prediction for mood annotation performs comparably to tag-based genres indicates that relying solely on audio in making predictions for novel tracks is a viable approach when human-generated semantic data are not available.

To further confirm the benefit of genre-adaptivity, ACTwg-SLPwg was applied to TE600 by first randomly rearranging the tag- and audio-based genre weights. It was assumed that if the high performance of ACTwg-SLPwg thus obtained would not degrade, the performance could be attributed to the benefit of ensemble modelling [63] and not to genre-adaptivity. The analysis showed that this is not the case. The genre randomisation degraded the prediction performance consistently. The average performance across mood scales dropped to $R^2 = 0.424, 0.400$ using tag- and audio-based genres respectively, and the performance difference was statistically significant at $p < 0.05$ for seven scales (tag-based genres) and for all scales (audio-based genres).

In summary, the results suggest that genre-adaptivity in both audio-based modelling and semantic computing is beneficial and that combining these two forms of genre adaptivity yields the highest performance improvements for music mood annotation.

7.3.3 Comparison of ACTwg-SLPwg and Genre-specific Models

If automatic music annotation is applied to a music collection representing one particular genre, one could ask whether a genre-specific model corresponding to the matching genre would be more appropriate than ACTwg-SLPwg. Such hypothesis was tested by comparing the prediction performance of ACTwg-SLPwg for the core affects separately on subsets of TE600 associated to the six genre clusters. TE600 was split for this purpose using tag-based genres. Table 10 showsthe results obtained with SLP for each subset. The genre-spcific ACTwg-SLPwg sub-model corresponding to the genre of the subset (using no genre-weighting, see (10)), and ACTwg-SLPwg with audio-based genres.

In this analysis, ACTwg-SLPwg yielded consistently higher performance than the genre-specific models and SLP, with only few exceptions: Valence/Electronic, Arousal/Metal and Tension/Electronic. Compared to general SLP, the genre-specific models were more successful at predicting Valence, which provides further evidence that genre-specific aspects need to be taken into account when modelling the Valence dimension. On the other hand, the results for Arousal showed an opposite pattern. These results corroborate the findings of [34], where audio-based genre-specific models of Arousal generalised better across genres than those of Valence.

TABLE 10
Prediction performance of SLP, genre-specific model and ACTwg-SLPwg separately for tracks from different genres.

		SLP	Genre-specific	ACTwg -SLPwg
Valence	Metal	0.387 0.020	0.407 0.011	**0.421*** 0.008
	Folk	0.199 0.019	0.127 0.037	**0.267*** 0.006
	Electronic	0.239 0.022	**0.339*** 0.009	0.316* 0.014
	Jazz	0.267 0.024	0.305 0.021	**0.360*** 0.012
	Rock	0.311 0.019	0.351* 0.003	**0.378*** 0.004
	Pop	0.225 0.020	0.299* 0.009	**0.306*** 0.006
Arousal	Metal	**0.720** 0.006	0.584 0.011	0.713 0.008
	Folk	0.703 0.006	0.674 0.006	**0.715*** 0.003
	Electronic	0.735 0.004	0.727 0.003	**0.748*** 0.003
	Jazz	0.671 0.009	0.642 0.008	**0.686** 0.008
	Rock	0.723 0.005	0.707 0.006	**0.733*** 0.006
	Pop	0.713 0.004	0.716 0.002	**0.723*** 0.004
Tension	Metal	0.541 0.014	0.499 0.022	**0.571** 0.004
	Folk	0.379 0.022	0.321 0.030	**0.415*** 0.007
	Electronic	0.372 0.019	**0.424*** 0.030	0.415* 0.007
	Jazz	0.358 0.020	0.336 0.011	**0.399*** 0.007
	Rock	0.473 0.018	0.469 0.005	**0.505*** 0.004
	Pop	0.426 0.025	0.443 0.009	**0.465*** 0.010

* $p < .05$ for improvement over SLP.

Overall, the genre-adaptive technique was clearly more successful than the genre-specific models. Since genre-specific models rely on training data from one genre, the models may suffer from low variance in the mood content, and might not therefore tap into more general relationships between audio features and mood, only attainable from collections of tracks spanning multiple genres.

7.3.4 The Impact of the Number of Genres

To explore the role of the number of genre clusters on the performance of ACTwg-SLPwg with audio-based genres, analysis was carried out using the genre clusterings with 2-16 genres (cf. Table 3). The results shown in Fig. 6, demonstrates that ACTwg-SLPwg performance is not overly sensitive to the number of genres. Performance consistently remains at a higher level than that of SLP on all genre clusterings. The optimal performance was found for all of the core affects at $K = 6$, which may possibly be attributed to the fact that TE600 is balanced according to the corresponding genres.

8 Conclusion

The present study examined how genre information can be incorporated into music mood prediction using genre-adaptive semantic computing and genre-adaptive audio-based modelling. As the general baseline technique, SLP performed favourably when compared to a state-of-the-art auto-tagging method. The comparison with genre-adaptive

TABLE 9
Performance of the genre-split/combine techniques with genres inferred from tags and audio. Performance improvements over the ACT (ACTwg) or SLP (other techniques) are highlighted.

	Tag-based genres				Audio-based genres			
	ACTwg	ACT -SLPwg	ACTwg -SLP	ACTwg -SLPwg	ACTwg	ACT -SLPwg	ACTwg -SLP	ACTwg -SLPwg
Valence	**0.457** 0.005 *	**0.434** 0.017†*	**0.406** 0.004†*	**0.457** 0.004 *	**0.456** 0.004†*	**0.397** 0.018†*	**0.406** 0.004†*	**0.431** 0.004 *
Arousal	**0.488** 0.003†	0.722 0.006	**0.741** 0.002†*	**0.732** 0.003	**0.489** 0.004†	0.725 0.006†	**0.741** 0.002 *	**0.741** 0.005 *
Tension	**0.397** 0.004†*	**0.505** 0.017	**0.513** 0.005†*	**0.518** 0.004 *	**0.400** 0.005†*	**0.497** 0.017†	**0.513** 0.005 *	**0.520** 0.005 *
Atmospheric	0.391 0.033†	0.689 0.013†	**0.700** 0.011†	0.631 0.035 *	**0.423** 0.027†	**0.699** 0.014	**0.703** 0.012	0.689 0.024
Happy	**0.431** 0.004†*	**0.367** 0.032 *	**0.358** 0.014†*	**0.389** 0.006 *	**0.434** 0.008†*	**0.331** 0.031†	**0.362** 0.017 *	**0.369** 0.012 *
Dark	**0.366** 0.027†	**0.300** 0.025 *	0.197 0.026†*	**0.268** 0.041	**0.381** 0.025†	**0.275** 0.021	0.208 0.031 *	**0.270** 0.037
Sad	**0.310** 0.005†*	0.288 0.012†	**0.333** 0.008 *	**0.330** 0.005 *	**0.317** 0.007†*	0.291 0.011†	**0.340** 0.007 *	**0.338** 0.006 *
Angry	**0.553** 0.005†*	**0.643** 0.016 *	**0.603** 0.008†*	**0.643** 0.006 *	**0.552** 0.007†*	**0.629** 0.017 *	**0.602** 0.007†	**0.639** 0.004 *
Sensual	0.357 0.024†*	**0.546** 0.009†	0.480 0.047 *	0.517 0.045	0.384 0.015†	**0.545** 0.011	0.511 0.024†*	**0.546** 0.021
Sentimental	**0.255** 0.008†	0.282 0.027	**0.334** 0.015	**0.338** 0.017	**0.283** 0.013†*	0.289 0.026†	**0.369** 0.019	**0.377** 0.017 *
Average	0.401	0.478	0.466	0.482	0.412	0.468	0.476	0.492

† $p < .05$ for performance difference between the ACTwg-SLPwg.

* $p < .05$ for performance difference between the ACT (ACTwg) or the SLP (other techniques).

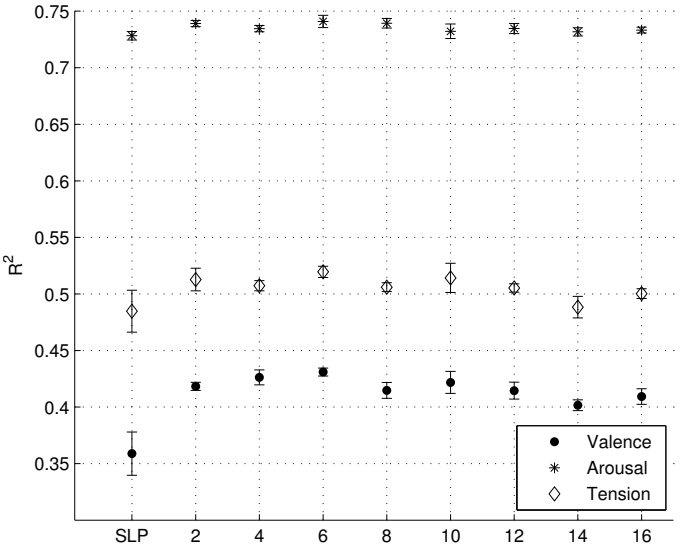


Fig. 6. The Median ± MAD performance of ACTwg-SLPwg using genre clusterings with 2-16 genres. The SLP performance is shown for comparison.

models showed that taking into account the genre information in mood annotation yields consistent improvements. The highest performing novel technique ACTwg-SLPwg models both the semantic mood space and audio-to-mood relationship in a genre-adaptive manner. Moreover, audio-based genre inference for a novel track performed favourably compared to tag-based inference, which has positive implications for applying the models to large unannotated datasets.

The study also offered survey results and analytical insights into inferring concise music genre representations based on a large set of genre tags. Moreover, the study demonstrated that semantic modelling of mood space based on music-specific social tag data is not surpassed by non-music-specific normative data obtained from a controlled laboratory survey.

The proposed techniques could be applied to other tasks, such as object recognition from images, video auto-tagging, or multimedia retrieval, where context-adaptive semantic modelling combined with context-adaptive content-based prediction could be beneficial.

ACKNOWLEDGMENTS

The work of Pasi Saari is funded by the Academy of Finland (The Finnish Centre of Excellence in Interdisciplinary Music Research). The work of György Fazekas and Mark Sandler is funded by the EPSRC programme grant "Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption" (FAST-IMPACt) EP/L019981/1.

REFERENCES

[1] T. Schäfer and P. Sedlmeier, "From the functions of music to music preference," *Psychology of Music*, vol. 37, no. 3, pp. 279–300, 2009.

[2] M. R. Zentner, D. Grandjean, and K. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[3] X. Hu and J. S. Downie, "Exploring mood metadata: relationships with genre, artist and usage metadata," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[4] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models," in *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2012, pp. 492–507.

[5] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proceedings of the 11th International Conference of Music Information Retrieval (ISMIR)*. Citeseer, 2010, pp. 255–266.

[6] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content," *IEEE Signal Processing Magazine : Special Issue on Semantic Retrieval of Multimedia*, vol. 23, pp. 133–141, 2006.

[7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[8] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[9] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 383–395, 2009.

[10] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music." *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[11] P. N. Juslin and J. A. Sloboda, "Introduction: Aims, organization, and terminology," in *Handbook of Music and Emotion: Theory, Research, Applications*. Boston, MA: Oxford University Press, 2010, pp. 3–14.

[12] T. Bertin-Mahieux, D. Eck, F. Maillet., and P. Lamere, "Autotagger: A model for predicting social tags from acoustic features on large music databases," *Journal of New Music Research*, vol. 37, no. 2, pp. 115–135, 2008.

[13] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[14] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, Feb. 2008.

[15] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 6, pp. 1802 –1812, aug. 2011.

[16] X. N. Lam, T. Vu, T. D. Le, and A. D. Duong, "Addressing cold-start problem in recommendation systems," in *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, 2008, pp. 208–211.

[17] M. R. Zentner and T. Eerola, "Self-report measures and models," in *Handbook of Music and Emotion: Theory, Research, Applications*, P. N. Juslin and J. A. Sloboda, Eds. Boston, MA: Oxford University Press, 2010, pp. 187–221.

[18] E. Law and L. Von Ahn, "Input-agreement: a new mechanism for collecting data using human computation games," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 1197–1206.

[19] Y. E. Kim, E. Schmidt, and L. Emelle, "Moodswings: A collaborative game for music mood label collection," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, 2008, pp. 231–236.

[20] J. C. Wang, Y. H. Yang, K. Chang, H. M. Wang, and S. K. Jeng, "Exploring the relationship between categorical and dimensional emotion semantics of music," in *Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*. ACM, 2012, pp. 63–68.

[21] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler, "Semantic models of mood expressed by music: Comparison between crowd-sourced and curated editorial annotations," in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2013, pp. 1–6.

[22] C. Laurier, M. Sordo, J. Serra, and P. Herrera, "Music mood representations from social tags," in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 381–86.

[23] P. Saari and T. Eerola, "Semantic computing of moods based on tags in social media of music," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2548–2560, 2014.

[24] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, April 2006.

[25] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[26] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, pp. 169–200, 1992.

[27] M. Levy and M. Sandler, "A semantic space for music derived from social tags," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[28] E. Law, B. Settles, and T. Mitchell, "Learning to tag from open vocabulary labels," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, J. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds. Springer Berlin Heidelberg, 2010, vol. 6322, pp. 211–226.

[29] P. Saari, T. Eerola, G. Fazekas, and M. Sandler, "Using semantic layer projection for enhancing music mood prediction with audio features," in *Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013)*, 2013, pp. 722–728.

[30] P. Saari, T. Eerola, G. Fazekas, M. Barthet, O. Lartillot, and M. Sandler, "The role of audio and tags in music mood prediction: A study using semantic layer projection," in *Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR)*, 2013.

[31] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 576–588, 2010.

[32] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendation: State of the art and challenges," *Computer Science Review*, vol. 6, no. 2, pp. 89–119, 2012.

[33] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, pp. 40:1–40:30, 2012.

[34] T. Eerola, "Are the emotions expressed in music genre-specific? an audio-based evaluation of datasets spanning classical, film, pop and mixed genres," *Journal of New Music Research*, vol. 40, no. 4, pp. 349–366, 2011.

[35] B. Schuller, H. Hage, D. Schuller, and G. Rigoll, "'Mister D.J., cheer me up!': Musical and textual features for automatic mood classification," *Journal of New Music Research*, vol. 39, no. 1, pp. 13–34, 2010.

[36] Y.-C. Lin, Y.-H. Yang, and H.-H. Chen, "Exploiting genre for music emotion classification," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2009, pp. 618–621.

[37] Y.-C. Lin, Y.-H. Yang, and H. H. Chen, "Exploiting online music tags for music emotion classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 7, no. 1, pp. 26:1–26:16, 2011.

[38] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant." *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 805–819, 1999.

[39] U. Schimmack and A. Grob, "Dimensional models of core affect: A quantitative comparison by means of structural equation modeling," *European Journal of Personality*, vol. 14, no. 4, pp. 325–345, 2000.

[40] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford University Press, New York, USA, 1989.

[41] T. Li and M. Ogihara, "Detecting emotion in music," in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*, 2003, pp. 239–240.

[42] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*. ACM, 2010, pp. 267–274.

[43] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 621–626.

[44] S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[45] J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[46] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[47] M. Delsing, T. ter Bogt, R. Engels, and W. Meeus, "Adolescents music preferences and personality characteristics," *European Journal of Personality*, vol. 22, no. 2, pp. 109–130, 2008.

[48] P. J. Rentfrow, L. R. Goldberg, and D. J. Levitin, "The structure of musical preferences: A five-factor model." *Journal of Personality and Social Psychology*, vol. 100, no. 6, pp. 1139–1157, 2011.

[49] M. Sordo, Ò. Celma, M. Blech, and E. Guaus, "The quest for musical genres: Do the experts and the wisdom of crowds agree?" in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, 2008.

[50] S. R. Ness, A. Theocharis, G. Tzanetakis, and L. G. Martins, "Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs," in *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 2009, pp. 705–708.

[51] Y.-H. Yang, Y.-C. Lin, A. Lee, and H. H. Chen, "Improving musical concept detection by ordinal regression and context fusion." in *Proceedings of the 10th International Conference of Music Information Retrieval (ISMIR)*, 2009, pp. 147–152.

[52] R. Miotto and G. Lanckriet, "A generative context model for semantic music annotation and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1096–1108, 2012.

[53] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx)*, September 2007.
[54] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*. Oxford University Press, 2004, vol. 3.
[55] K. R. Scherer, "Emotion as a multicomponent process: A model and some cross-cultural data," in *Review of Personality and Social Psychology*. Beverly Hills: CA: Sage, 1984, vol. 5, pp. 37–63.
[56] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models and stimuli," *Music Perception*, vol. 30, no. 3, pp. 307–340, 2013.
[57] V. Warriner, Amy Bethand Kuperman and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, pp. 1–17, 2013.
[58] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
[59] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*. Springer Berlin Heidelberg, 2004, pp. 39–50.
[60] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 14. Cambridge, MA: MIT Press, 2001, pp. 849–856.
[61] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, april 1979.
[62] B. Mirkin, *Mathematical Classification and Clustering*. Kluwer Academic Press, Dordrecht, 1996.
[63] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

Pasi Saari received MSc degree in computer science in 2008 and MA degree in musicology in 2010 from the University of Jyvaskyla, Finland. In 2010-2014 he worked as a doctoral student at the Finnish Centre of Excellence in Interdisciplinary Music Research within the same institution. Prior to finishing the studies, Saari worked as a Principal Researcher at Nokia Technologies, Tampere, Finland. In the early 2015, he obtained a PhD degree with a thesis titled Music Mood Annotation Using Semantic Computing and Machine Learning. He is currently working as a Postdoctoral Research Associate at the Department of Music at the Durham University, UK. His research interests lie within the field of Music Information Retrieval, in particular in the computational modelling of musical emotions and the contexts of music listening.

György Fazekas is a lecturer at Queen Mary University of London, working at the Centre for Digital Music (C4DM), School of Electronic Engineering and Computer Science. He received his BSc degree at Kando Kalman College of Electrical Engineering, Obuda University, Faculty of Electrical Engineering. He received an MSc and PhD degree at Queen Mary University of London, UK in 2012. His thesis titled Semantic Audio Analysis–Utilities and Applications explores novel applications of semantic audio analysis, Semantic Web technologies and ontology-based information management. His research interests include the development of semantic audio technologies and their application to creative music production. He is involved in collaborative research projects and he is a member of the IEEE, AES, ACM and BCS.

Tuomas Eerola received his MA and Ph. D degrees from the University of Jyvaskyla, Finland, in 1997 and 2003. He is a Professor of Music Cognition at the Durham University, UK. His research interest lies within the field of music cognition and music psychology, including musical similarity, melodic expectations, perception of rhythm and timbre, and induction and perception of emotions. He is on the editorial boards of Psychology of Music, and Frontiers in Digital Humanities and is consulting editor for Musicae Scientiae and member of the European Cognitive Sciences of Music (ESCOM) and the Society for Education, Music and Psychology Research (SEMPRE).

Mathieu Barthet is a Lecturer in Digital Media at Queen Mary University of London. He received the M.Sc degree in Acoustics from Aix-Marseille II University and Ecole Centrale Marseille (France) in 2004. He was awarded a Ph.D. from Aix-Marseille II University and CNRS-Laboratory of Mechanics and Acoustics (CNRS-LMA) in 2008 ("From performer to listener: an acoustical and perceptual analysis of musical timbre"). From 2009 to 2014 he was a Postdoctoral Researcher with the Centre for Digital Music at Queen Mary University of London where he worked on projects in collaboration with the BBC, the British Library, and I Like Music. His research interests include music informatics, affective computing, human computer interaction, music perception, computational musicology, and big music data. In 2012 he was general chair of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR) conference "Music and Emotions". He is a guitarist and regularly plays in Pop/Rock and Jazz ensembles.

Olivier Lartillot is a researcher in computational music analysis at the Department for Architecture, Design and Media Technology, Aalborg University, Denmark. Formerly at the Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyvaskyla, he designed MIRtoolbox, a referential tool for music feature extraction from audio. He also works on symbolic music analysis, notably on sequential pattern mining. In the context of his 5-year Academy of Finland research fellowship, he conceived the MiningSuite, an analytical framework that combines audio and symbolic research. He continues his work as part of a collaborative European project called Learning to Create (Lrn2Cr8).

Mark Sandler (SM98) was born in 1955. He received the B.Sc. and Ph.D. degrees from the University of Essex, Essex, U.K., in 1978 and 1984, respectively. He is a Professor of Signal Processing at Queen Mary University of London, London, U.K., and Head of the School of Electronic Engineering and Computer Science. He has published over 350 papers in journals and conferences. Prof. Sandler is a Fellow of the Institute of Electronic Engineers (IEE) and a Fellow of the Audio Engineering Society. He is a two-time recipient of the IEE A. H. Reeves Premium Prize.